Cassandra Crossing/ Archivismi: archiviamo Cassandra, parte terza

(566)—E' tempo di concludere; parte il mass uploading di Cassandra Crossing!

Cassandra Crossing/ Archivismi: archiviamo Cassandra, parte terza
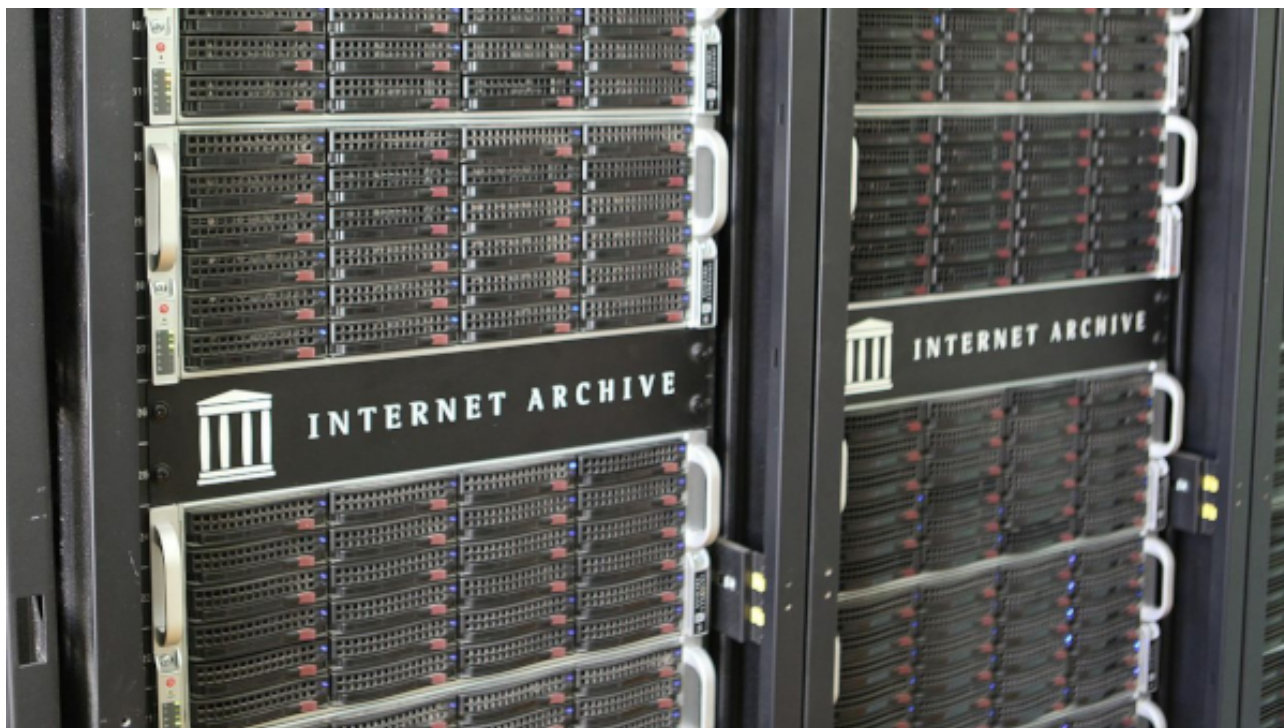
Figure 1:

(566)—E' tempo di concludere; parte il mass uploading di Cassandra Crossing!

5 gennaio 2024—Nelle precedenti puntate di *Archivismi* abbiamo raccontato come funziona, a grandi linee, una archiviazione "*vera*" su Internet Archive. "*Vera*" perché non si tratta di caricare una directory di file, ma di creare veri oggetti archivistici, corredati di tutti i file ed i metadati necessari per definire l'oggetto, e renderlo utile e fruibile. Ed i metadati, credeteci o no, sono di gran lunga la cosa più difficile e più utile.

Quindi, innanzitutto, per archiviare la nostra rubrica preferita, è stato necessario chiedersi *cosa* archiviare, oltre al classico PDF. La scelta è stata quella di aggiungere un file HTML entrocontenuto ed un file in formato MARKDOWN, quest'ultimo utile per ulteriori elaborazioni che fossero necessarie. Alcuni articoli parlavano inoltre di libri o pubblicazioni libere, ed in questi pochi casi anche il pdf della pubblicazione è stato inserito nell'oggetto.

Bene, detto questo, è stato necessario crearli, questi benedetti 1686 file. I file markdown, html e e pdf sono stati generati in completo automatismo a partire dai file html degli articoli esportati da Medium.com, grazie agli strumenti preparati nelle puntate precedenti che erano pronti all'uso, elaborando i dati di input esportati da Medium.com. Tutto semplice, quindi?

Ovviamente no. In questi appunti di viaggio, la vostra profetessa preferita vi racconterà le ulteriori peripezie incontrate nel suo viaggio.

Uno: i dati da Medium.com contenevano ancora degli errori. La tipologia più comune e più dolorosa era l'errata costruzione del nome del file, creato rilevando automaticamente il numero dell'articolo. Questo per due ragioni principali. La prima è che alcuni articoli erano semplicemente numerati in maniera errata. La seconda è che i file contenevano si il numero dell'articolo, ma non solo nel testo, anche nella intestazione creata automaticamente da Medium.com. Intestazione che una volta creata non veniva poi più aggiornata; indovinate da dove veniva preso il numero dell'articolo?

Due: la creazione del foglio elettronico, avendo i file ben creati e rinominati è stata semplice. Aver conservato ogni run di upload in un nuovo foglio è stato utilissimo per localizzare gli errori e ritornare sui propri passi. Anche conservare il log delle esecuzioni di *ia* è stato utilissimo per estrarre gli errori.

Tre: *aggiustando* la numerazione degli articoli in alcuni casi si è persa la corrispondenza tra nome dei file ed identificativo dell'oggetto. Infatti, mentre i file ed i metadati si possono modificare, aggiungere e cancellare, non è possibile modificare l'identificativo del l'oggetto, una volta creato. E quando si lancia nuovamente la procedura di generazione file, se cambia la numerazione cambiano anche alcuni nomi di file. Per generare i successivi fogli per il caricamento è stato necessario tenere conto di questo, e operare esaustive verifiche di *allineamento* tra identificatori e nomi dei file. Certo, la tentazione di correggere tutto e rilanciare daccapo le procedure era forte. Ma l'automazione totale non è il fine, ma solo un mezzo. Risparmiare tempo, **facendo comunque le cose per bene,** è il vero fine.

Quattro: il primo bulk upload del solo file PDF è stato fatto per 10 oggetti. Si è poi atteso che le varie alchimie automatiche di Internet Archive si compissero, e si è esaminato attentamente il risultato. A livello di metadati questo ha portato a modificare le scelte per renderli più utili.

Cinque: Si è poi fatto il bulk upload dei rimanenti 552 pdf, creando così tutti gli oggetti. Gli oggetti, ed in particolare gli identificatori, in tutte le successive operazioni che abbiamo fatto non sono mai variati. Durante questo primo vero bulk upload si sono generati messaggi di errore di *mancata creazione,* perché l'operazione in corso era stata identificata come spam, come questo

error uploading 186_Cassandra-Crossing—L-Internet-senza-Rete.pdf: Please reduce your request rate.—Your upload of 186_Cassandra-Crossing—L-Internet-senza-Rete from username pippo@pluto.paperino appears to be spam. If you believe this is a mistake, contact info@archive.org and include this entire message in your email.

Detto fatto, ho contattato via email l'help desk che, forse perché sono un utente di vecchia data nonché *donatore* regolare, in poche ore mi ha tolto qualche evidente limitazione antispam. I successivi inserimenti non hanno più dato nessun problema.

Sei: Sono stati eseguiti due ulteriori bulk upload separati, uno per i file markdown ed uno per gli html. Sono state necessarie solo due colonne nei fogli elettronici; identificatore e file. I metadati sono stati assegnati al momento della creazione dell'oggetto, quindi del primo bulk upload. Se dovessero essere cambiati in massa, sarà necessario effettuare "*bulk correction*".

Sette: si sono appunto editati i metadati in bulk, inserendo la descrizione (presa dal sottotitolo) e la data di pubblicazione. Ambedue queste colonne di dati sono state generate con una versione modificata della procedura già vista, partendo dai file markdown, estraendo il campo con una regular expression, aggiungendo, ripulendo e correggendo i campi mancanti od errati a mano, e poi copiando i range giusti nel foglio elettronico per il bulk upload. Malgrado le "*standardizzazioni*" delle precedenti fasi di redazione e manipolazione dei file degli articoli, per sistemare le discrepanze c'è voluta più di mezza giornata.

Otto: E qualche altra ora c'è voluta per esaminare sul sito di Internet Archive l'elenco degli articoli ordinati per data e vedere che dentro ci fosse quello che ci deve essere. Anche qui qualche piccolo errore è emerso, ma solo di data. Solo in un caso i titoli e le date erano ambedue invertiti, ma per fortuna anche questi sono metadati, quindi facilmente correggibili. Ma è stata anche una soddisfazione ripercorrere venti anni di lavoro in poche ore!

Ed anche per oggi è tutto, perché il lavoro di revisione è davvero stancante. Le conclusioni ed i commenti li riserviamo per la prossima e finale puntata di questa prima campagna di "*Archivismi*".

Scrivere a Cassandra—Twitter—Mastodon
Videorubrica "Quattro chiacchiere con Cassandra" tempo
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

By Marco A. L. Calamari on January 5, 2024.

Canonical link

Exported from Medium on August 27, 2025.